*D646–D652* *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 12 November 2011*
*doi:10.1093/nar/gkr990*

# The UCSC Archaeal Genome Browser: 2012 update

**Patricia P. Chan, Andrew D. Holmes, Andrew M. Smith, Danny Tran and Todd M. Lowe***

Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, SOE-2, Santa Cruz, CA 95064, USA

Received October 1, 2011; Accepted October 17, 2011

## ABSTRACT

**The UCSC Archaeal Genome Browser (http://archaea.ucsc.edu) offers a graphical web-based resource for exploration and discovery within archaeal and other selected microbial genomes. By bringing together existing gene annotations, gene expression data, multiple-genome alignments, pre-computed sequence comparisons and other specialized analysis tracks, the genome browser is a powerful aggregator of varied genomic information. The genome browser environment maintains the current look-and-feel of the vertebrate UCSC Genome Browser, but also integrates archaeal and bacterial-specific tracks with a few graphic display enhancements. The browser currently contains 115 archaeal genomes, plus 31 genomes of viruses known to infect archaea. Some of the recently developed or enhanced tracks visualize data from published high-throughput RNA-sequencing studies, the NCBI Conserved Domain Database, sequences from pre-genome sequencing studies, predicted gene boundaries from three different protein gene prediction algorithms, tRNAscan-SE gene predictions with RNA secondary structures and CRISPR locus predictions. We have also developed a companion resource, the Archaeal COG Browser, to provide better search and display of arCOG gene function classifications, including their phylogenetic distribution among available archaeal genomes.**

## INTRODUCTION

The feature-rich UCSC Genome Browser (1,2), created originally to annotate the human genome, has become an established graphical web resource for analyzing higher eukaryotes. Its extensible architecture and the relatively small size of microbial genomes enabled us to develop the Archaeal Genome Browser with modest resources 6 years ago (3). The initial set of 26 archaeal genomes contained basic annotation tracks including G/C (guanine/cytosine) content, protein gene predictions from NCBI RefSeq (4) and the Comprehensive Microbial Resource (5), as well as known and predicted non-coding RNA genes. Additional details for protein gene annotations derived from Pfam domains (6), COG groups (7), KEGG pathway information (8) and ModBase structure predictions (9) were also integrated. Computational predictions of promoters and Shine–Dalgarno motifs (Chan and Lowe, manuscript in preparation), microarray gene expression studies (10) and multiple-genome alignments (11,12) further enriched the content of the genome browser with diverse gene information and enabled comparative genomic analysis.

With the increased availability of complete archaeal genomes, we expand our database to include 115 archaeal genomes, the genomes of 31 viruses known to infect archaea and more than 250 bacterial genomes. In addition, we develop new tracks to provide more diverse or detailed information based on NCBI conserved protein domains, CRISPR predictions and paralogs within genome. The newly implemented arCOGs Browser presents the Archaeal Clusters of Orthologous Genes (13), and further enables users to classify genes and their functions. To analyze gene expression patterns and discover novel transcripts, we include 19 microarray and two RNA sequencing (RNA-seq) data sets in six different archaeal species; we expect many more RNA-seq data sets in the near future as these functional genomics studies become more common. Together with new functionality in the UCSC Genome Browser, the information combined from diverse sources provides a valuable resource for the archaeal research community.

## NEW BROWSER FEATURES

### Updated entry portal

We have redesigned the home page of the Archaeal Genome Browser (archaea.ucsc.edu) for better presentation of database resources and easier accessibility to the rapidly growing genome browser collection. From the entry portal, a series of 'tabs' provides direct access to

*To whom correspondence should be addressed. Tel: +1 831 459 1511; Fax: +1 831 459 4829; Email: lowe@soe.ucsc.edu

supporting information or resources. A 'News' section highlights newly added genomes, new functional genomics data sets and improved browser features. To help researchers use the genome browser more effectively, a 'Tutorials' section now contains short video guides and slide presentations, covering basic navigation, core tracks, browser configuration and advanced data search/extraction using the powerful 'table browser' database interface (14). Other new sections accessible from the home page include links to the new arCOGs browser, summary information about all available functional genomics data sets, a link to 'Gene-Pub' (described below) and a 'Resources' page with a collection of links to general information about archaea, archaeal research labs and other genome analysis resources.

### Genome access and description

The standard method of launching the genome browsers has been the selection of genomes based on major clades at the genome gateway pages. To provide easier access, we have added a quick search box on the home page of the genome browser website. When users enter a partial genome name, the search box suggests matching available genomes, with links to the genome browser and species information/gateway page. Users may also peruse the full list of species from the 'Genomes' tab on the home page.

At each species' information/gateway page, researchers can view basic information including genome size, a listing of all associated chromosomes and plasmids, the number of predicted genes, G/C content, taxonomy and a direct link to the source sequence RefSeq entry at NCBI. Other species-specific information relating to the organism's habitat and physiology can be viewed via abstracts and PubMed links to the primary literature which detail the species' isolation and genome sequencing, if available (Figure 1A). Finally, the species' phylogeny among closely related species can be viewed in a phylogram that is computed from a multiple-genome alignment (12,15); the alignment itself is viewable within the 'Conservation' track of the species' genome browser.

### Feature set browsing and information tooltip

The relatively small number of genes in each microbial genome enables us to provide a feature-list browsing window for basic annotation tracks including NCBI RefSeq (4) coding and non-coding genes, tRNAscan-SE (16) gene predictions, Pfam protein domains (6), Rfam non-coding RNA predictions (17) and CRISPR predictions (18,19) (Figure 1A). By selecting an available feature set from the genome information/gateway page, a list of the genes or features with the name, description and position in the genome can be displayed on the left frame while the genome browser is displayed in the right frame for the selected feature (Figure 1B). In this way, users can browse, select and inspect the genomic context of many features in a track rapidly, without shifting between windows.

For ease of navigation, we have introduced an information tooltip function within the browser display. When users move the mouse over a track item such as a gene in the NCBI RefSeq track, the feature name and brief description will be shown in a popup box without launching the item description page (Figure 1B).

### New features from UCSC Genome Browser

By using the code base developed for the eukaryote-specific UCSC Genome Browser, we have been able to adopt its regularly improving functionality into the Archaeal Genome Browser. Users can now move the genome display region by dragging the browser image left and right, and zoom in to a desired region by holding the shift key while dragging over it. Annotation tracks can also be reordered by dragging them up or down in the browser window. To enable the display of very large data sets and sequence alignments from high-throughput sequencing results, the genome browser group has introduced BigBed and BigWig (20) file formats and included the support of the Binary Alignment/Map (BAM) (21) file format. A new data hub feature also allows researchers to create custom BigBed, BigWig or BAM tracks in groups (composite tracks) and host the data at local web servers. To facilitate user-directed genome analyses with a variety of tools, researchers can export data directly from the UCSC table browser to Galaxy (22), an external, interactive platform for computational biological research. Users may also access the table browser functions of the Archaeal Genome Browser directly on the Galaxy website, and can visualize the results of new Galaxy data analyses within our genome browsers by exporting custom tracks from Galaxy.

## NEW DATA

### Genome assemblies

Since our initial publication in 2006, we have continued to add complete archaeal genomes as they become available, as well as draft genomes on request. To date, the Archaeal Genome Browser contains 115 archaeal genomes; broken down into subdomains, these include 35 crenarchaea, 75 euryarchaea, 2 thaumarchaea, 1 nanoarchaeon, 1 korarchaeon and 1 *Caldiarchaeum*. We have also provided 277 bacterial genomes and 31 archaeal viral genomes for comparative studies. In collaboration with other research labs, we have included seven mycobacteriophages and the human malaria parasite *Plasmodium falciparum*.

### Browser tracks

Besides the gene annotations from NCBI RefSeq (4) and Comprehensive Microbial Resource (5), we have created new tracks in the standard collection for archaeal genomes that further diversify the sources of genome annotation (Figure 2). These include gene predictions from the Integrated Microbial Genomes system (23); individual gene sequences that appear as independent entries in GenBank (generally from pre-genome age gene locus studies) (24); and protein-coding gene predictions using Glimmer (25), GeneMark (26) and Prodigal (27).

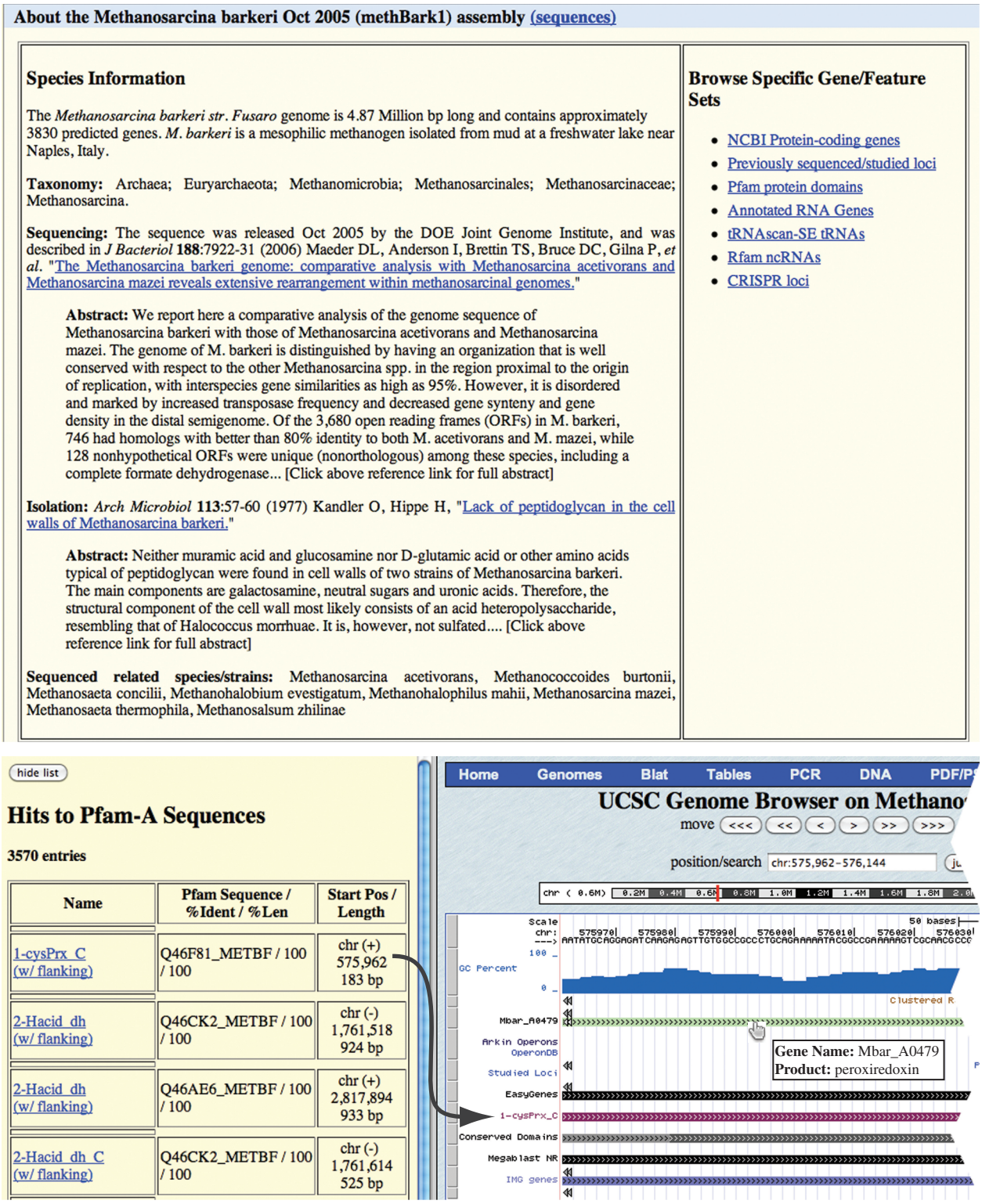


**Figure 1.** Genome description page layout and feature browsing. (**A**) An example of a genome information/gateway page that includes genome size, the number of predicted genes, taxonomy, links to publications detailing the species' isolation and genome sequencing, and links to feature sets available for indexed browsing. (**B**) The Pfam (6) domains within the genome are displayed in the left frame after clicking on the 'Pfam protein domains' link described in Figure 1A. The right frame displays the genomic region of the feature selected from the left frame; in this example, the selected 1-cysPrx_C Pfam domain is displayed within the genome browser.


**Figure 2.** Browser tracks available for a sample genome, *P. furiosus*. (A) GC Percent graph showing the G+C content percentage within a 20-nt sliding window. (B) CRISPR predictions (18,19). (C) NCBI RefSeq (4) protein-coding gene annotations with track elements color-coded to indicate COG functional category (38). (D) Comprehensive Microbial Resource gene annotations (5). (E) Integrated Microbial Genomes gene annotations (23). (F) Protein-coding gene predictions using GeneMark (26), Glimmer (25) and Prodigal (27). (G) Operon predictions from MicrobesOnline (39) and OperonDB (40). (H) Independent gene entries at GenBank. (I) Genomic mappings of Pfam database (6) entries. (J) NCBI conserved domain database (28) search results using RPS-BLAST (29). (K) Non-coding RNA gene annotations from Rfam (17). (L) tRNA gene predictions by tRNAscan-SE (16). (M) NCBI RefSeq (4) non-coding RNA gene annotations. (N) Insertion sequence elements annotated by ISfinder (41).

(continued)

To better highlight conserved protein motifs, we have added genomic mappings of Pfam database (6) entries as a separate track, as well as matches to the NCBI conserved domain database (28) using RPS-BLAST (29). We also continue to update the protein sequence alignment tracks using BLASTP (29). The BLASTP alignment results enable us to identify potential paralogs within each genome. As the number of experimental characterization studies increases, valuable insights into gene function often are not integrated into existing genome annotation. We have developed Gene-Pub as a platform for establishing a link between published research and archaeal gene annotations. Researchers can access the Gene-Pub submission system through the website entry portal page, and provide their updated, experimentally supported annotation, with the corresponding publication(s) for any archaeal gene. The submitted information will be reviewed for accuracy and then appended to existing gene annotation.

Due to the existence of non-canonical tRNA introns and atypical tRNAs in archaea (30–34), the automated, usually dated annotations of these and other non-coding RNAs in RefSeq (4) may not be accurate. We therefore include a tRNA gene track predicted by an improved version of tRNAscan-SE (16,34) with links to the Genomic tRNA Database (35) for detailed information. The Rfam (17) track and CRISPR predictions (18,19), allow recognition of other known non-coding RNAs, giving the most complete view of non-coding RNAs in archaeal genomes.

From the Archaeal Genome Browser entry portal page, researchers can find a list of 19 available RNA sequencing and microarray experimental data sets that can be leveraged for gene expression analyses and novel transcript discovery. Links to Gene Expression Omnibus (GEO) (36), journal publications at PubMed, and the corresponding genome browsers are provided. Users may also retrieve the source RNA sequencing reads of *Pyrococcus furiosus* and *Sulfolobus solfataricus* by following links to the NCBI Sequence Read Archive (37). Within the genome browsers, we provide color-coded microarray tracks for displaying gene expression data. Furthermore, we have developed two separate sets of bar graphs to show the RNA-seq read coverage, as well as read-end density (Figure 2), enabling users to estimate the abundance level, length and boundaries of transcribed elements.

The increased number of complete microbial genomes in the past 5 years allows us to include multiple-genome nucleotide alignments among closely related species for most archaea and selected bacteria. For example, we provide a 14-way genome alignment in the browsers for Halobacteria. Enabling the visualization of annotation between species, the new 'aligned features' track can be used to identify orthologous genes in closely related genomes based on nucleotide sequence alignments. On the protein level, BLASTX tracks allow detection of potentially missed ORFs in intergenic regions, based on BLASTX comparison of unannotated nucleotide sequences to the protein databases.

## Archaeal COG Browser

Clusters of Orthologous Groups of proteins (COGs) have been widely used for evolutionary gene classifications, focusing primarily on bacterial and eukaryotic systems (38). To extend this work, the Evolutionary Genomics Research Group at NCBI developed the Archaeal Clusters of Orthologous Genes (arCOGs) that classify genes and provide improved functional annotation specific to archaeal genomes (13). With the input and support from the arCOGs team, we created the arCOGs Browser to provide better accessibility to this valuable resource. The current data set in the arCOGs browser represents the 2010 update, and will be updated periodically as releases are made available by the arCOGs team at NCBI. Researchers can search for gene loci and arCOG annotations across all archaeal genomes and view both the distribution and homolog count of any given arCOG or functional category (Figure 3). Each arCOG gene entry also links to the Archaeal Genome Browser for graphical viewing within its genome context. Within the genome browsers, arCOG annotations and links to the arCOG browser are reciprocally listed on the description page for each gene in the NCBI RefSeq track.

The arCOGs database and the new arCOGs browser address significant deficiencies in existing archaeal gene annotations. A large proportion of genes in archaeal genomes are annotated as hypothetical proteins due to the limited availability of biochemical and comparative genomic data when these genomes were initially sequenced. Even as new gene characterization studies are published and comparative genomic data multiply, annotation updates at Genbank using these information resources are seldom. For example, half of the genes in *Pyrobaculum aerophilum*, a crenarchaeon that was sequenced a decade ago, fall into this category. Using the arCOG data and browser, we found that more than 40% of these hypothetical proteins have an arCOG functional classification. Thus, integration of these new functional assignments within the Archaeal Genome Browsers represents an important advance for the research community.

**Figure 2.** Continued
(O) Palindromic transcription factor binding site predictions. (P) Promoter predictions on + strand using 1-nt sliding window of 16-nt BRE/TATA promoter motif scan. (Q) Poly-T motifs as possible transcription termination signals. (R) Ribosomal binding site predictions on + strand using 1-nt sliding window of 10-nt Shine–Dalgarno motif scan. (S) Small RNA sequencing data coverage by Michael Terns and colleagues (42). (T) Microarray expression data showing metabolism of elemental sulfur by Michael Adams and colleagues (43). (U) High similarity nucleotide alignments with other loci in the genome. (V) Paralogs within genome identified by BLASTP search (29). (W) Multiple sequence alignment similarity plot of closely related species using PhyloHMM (12,15). (X) Orthologous gene annotations in closely related genomes based on genome sequence alignments. (Y) Phylogenetic breakdown of BLASTP (29) protein similarities across all proteins within supported archaeal genomes.

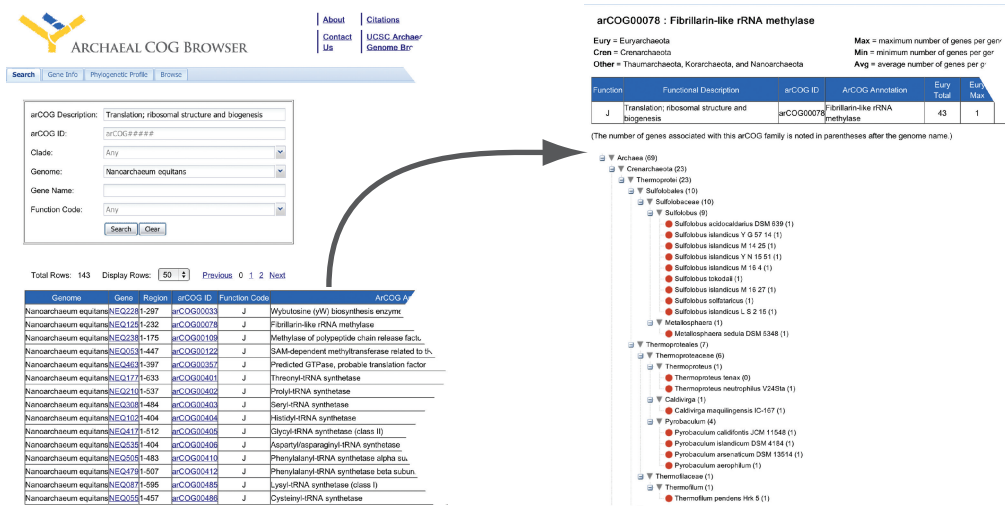


**Figure 3.** Using the Archaeal COG Browser. The left panel displays the main search interface with search results shown in the table below. Upon clicking on an arCOG link in the results table (in this example, the second entry arCOG00078), a new page displayed in the right panel gives the phylogenetic distribution of proteins within the arCOG (13).

## FUTURE DEVELOPMENT

We will continue to incorporate new archaeal genomes and update existing annotations in the Archaeal Genome Browser. Because the Archaeal Genome Browser shares the same code base as the actively developed UCSC Genome Browser, we expect to offer new feature updates regularly. With the expanded use of next-generation sequencing technologies, track updates to enhance accessibility and visualization of new functional data will be of growing importance. The Archaeal Genome Browser will continue to focus on providing complete genomes and publicly available annotations, although we encourage members of the research community to contribute genome-wide data sets and new analyses, as well as unpublished genomes still in the process of community-based annotation efforts.

## ACKNOWLEDGEMENTS

We thank Jim Kent and the Genome Bioinformatics Group of UC Santa Cruz for providing us with excellent assistance in developing the Archaeal Genome Browser. We are grateful to Lowe Laboratory members David Bernick, Aaron Cozen, Lauren Lui, Andrew Uzilov and Matthew Weirauch (University of Toronto) who helped develop annotation tracks. We thank Kira Makarova, Yuri Wolf and the arCOGs team in Evolutionary Genomics Research Group of NCBI for support and input in the development of the arCOGs Browser.

## FUNDING

Funding for open access charge: The National Science Foundation (DBI-0641061 and EF-0827055).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
2. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
3. Schneider,K.L., Pollard,K.S., Baertsch,R., Pohl,A. and Lowe,T.M. (2006) The UCSC Archaeal Genome Browser. *Nucleic Acids Res.*, **34**, D407–D410.
4. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
5. Davidsen,T., Beck,E., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K., White,O. *et al.* (2010) The comprehensive microbial resource. *Nucleic Acids Res.*, **38**, D340–D345.
6. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
7. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
8. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
9. Pieper,U., Eswar,N., Webb,B.M., Eramian,D., Kelly,L., Barkan,D.T., Carter,H., Mankoo,P., Karchin,R., Marti-Renom,M.A. *et al.* (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **37**, D347–D354.
10. Cozen,A.E., Weirauch,M.T., Pollard,K.S., Bernick,D.L., Stuart,J.M. and Lowe,T.M. (2009) Transcriptional map of respiratory versatility in the hyperthermophilic crenarchaeon Pyrobaculum aerophilum. *J. Bacteriol.*, **191**, 782–794.
11. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
12. Siepel,A. and Haussler,D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.

13. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct.*, **2**, 33.
14. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
15. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
16. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
17. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
18. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
19. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
20. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
22. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
23. Markowitz,V.M., Chen,I.M., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2010) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res.*, **38**, D382–D390.
24. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
25. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
26. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
27. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
28. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
30. Marck,C. and Grosjean,H. (2003) Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA*, **9**, 1516–1531.
31. Fujishima,K., Sugahara,J., Tomita,M. and Kanai,A. (2010) Large-scale tRNA intron transposition in the archaeal order Thermoproteales represents a novel mechanism of intron gain. *Mol. Biol. Evol.*, **27**, 2233–2243.
32. Randau,L., Munch,R., Hohn,M.J., Jahn,D. and Soll,D. (2005) Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature*, **433**, 537–541.
33. Fujishima,K., Sugahara,J., Kikuta,K., Hirano,R., Sato,A., Tomita,M. and Kanai,A. (2009) Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proc. Natl Acad. Sci. USA*, **106**, 2683–2687.
34. Chan,P.P., Cozen,A.E. and Lowe,T.M. (2011) Discovery of permuted and recently split transfer RNAs in Archaea. *Genome Biol.*, **12**, R38.
35. Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.
36. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
37. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
38. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
39. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
40. Pertea,M., Ayanbule,K., Smedinghoff,M. and Salzberg,S.L. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, **37**, D479–D482.
41. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
42. Hale,C.R., Zhao,P., Olson,S., Duff,M.O., Graveley,B.R., Wells,L., Terns,R.M. and Terns,M.P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, **139**, 945–956.
43. Schut,G.J., Bridger,S.L. and Adams,M.W. (2007) Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon Pyrococcus furiosus: characterization of a coenzyme A- dependent NAD(P)H sulfur oxidoreductase. *J. Bacteriol.*, **189**, 4431–4441.